IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11489

An Exploratory Study of ML Techniques in Football Match's Result Prediction

Dr. Kumud Kundu, Anurag Mishra, Ashish Kumar Singh, Apurav Sharma, Parth Arun

Department Of Computer Science and Engineering, Inderprastha Engineering College

Abstract—Machine learning is a subset of artificial intelligence (AI) that allows computers to learn and improvise on their own without having to be explicitly programmed. Machine learning deals with the creation of computer programs that can access data and learn on their own. Sports prediction is one of the rapidly-growing fields in good predictive accuracy since it involves a large sum of money in betting. The capability to apply algorithms and use that knowledge to try to forecast the outcome of future games based on this data is a particularly important aspect of machine learning in football. Sports match results can be difficult to forecast, with unexpected outcomes frequently occurring. Football is a good example since matches have a set length (as opposed to racket sports like tennis, where the game is played until one player wins). In this study, Machine Learning techniques are used to predict the winning team in the English Premier League (EPL). The goal is to predict a football match's full-time result (FTR) accurately, which determines the winning team. For training the data, we use algorithms like Support Vector Machines, XGBoost, and Logistic Regression, and the one with the highest and best accuracy is used to forecast the winning team. The data for previous seasons is obtained from [6].

Keywords—Football, Soccer Analytics, Prediction, Machine Learning, Support Vector Machine (SVM), XGBoost.

I. INTRODUCTION

Football is the most popular sport globally and is played by 250 million players in over 200 countries. Analytics has always been there in the field of sports even if we don't acknowledge it.

To be more precise, analytics in the field of football is the method of creating meaningful information and decisions that can be acted upon using soccer-related data. The data includes anything ranging from how many goals a team has scored to multiple factors like, distance covered by a player during the course of the match, or a number of passes played and how many out of those were accurate along with how many out of those created a chance for their team to score and so on.

In every soccer, league groups are formed and the teams play 2 matches with each alternative team in their league - one at their home structure and the other at the opponent's home stadium. Every such match has 3 doable outcomes the home side wins, the match ends in a draw, or the visiting team wins.

Given such a format, it's natural that there are many online fantasy leagues, betting agencies, and others who attempt to predict the end result of every match. during this project, an endeavor has been created to seek out the factors that have an effect on the outcome of a match and conjointly predict the results of any fixture by utilizing these factors.

The most important reason in the back of this venture is giving a correct dataset for football matches and predicting the winners in upcoming games and hence yielding efficient results. In this paper, we suggest a version of football prediction primarily based totally on FTR that is Full-Time Result(Our Class label) i.e. Home, Away, or Draw.

II. BACKGROUND

Outcomes from sports matches is difficult to predict, with surprises often doping up. Football specifically is a noteworthy example as matches have fixed length (as opposition racket sports like tennis, where the sport is played until a player wins).

However, because of the low-scoring nature of games (less than 3 goals per game on the average within the english premier league within the past 15 years) there's a random element linked to the quantity of goals scored during a match. There is a necessity to seek out if the appliance of machine learning can bring better and more insightful lead to soccer analytics. This makes match results an imperfect measure of a team’s performance and thus an incomplete metric on which to predict future results.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11489

III. LITERATURE SURVEY

Thamaraimanalan et al. [1] utilized the object detection within the football games depends upon the basic image processing techniques. They carried out training process using Convolutional Neural Network (CNN) classifier with the accuracy rate of 87.63% with the reduction in the sensitivity and the specificity range of 72.5 and 86.2%.

Yang [2] in their work proposed a linear support vector classifier (LSVC) to predict the outcome of the match based on the performance of the players. Model was validated with the results of the statistics of the AUC, F1 and prediction accuracy of the model were 0.8597, 0.6973 and 0.7965 respectively on the verification data.

Martinovic, J., Snásel, V., Ochodkova, E., Nolta, L., Wu, J. and Abraham, A. [3] in their work studied robot soccer game, as a part of standard applications of distributed system control in real time to predict strategy and perform game analysis.

Uhrín Matej, Šourek Gustav, Hubáček Ondřej, Železný Filip [4] investigated the two most prominent streams of betting investment strategies based on the views of the Modern Portfolio Theory and the Kelly criterion, together with a number of their popular modifications aimed at additional risk management in practice, where their original underlying mathematical assumptions do not hold.

Woo-Joo Lee, Hyo-Jin Jhang, Seung Hoe Choi [5] performed a study that aims to find variables that affect the winning rate of the football team before a match. Qualitative variables such as venue, match importance, performance, and atmosphere of both teams are suggested to predict the outcome Using Regression analysis.

Edward Wheatcroft [6] studied use of observed and predicted match statistics as inputs to forecasts for the outcomes of football matches. It is shown that, were it possible to know the match statistics in advance, highly informative forecasts of the match outcome could be made.

Rahul Baboota, Harleen Kaur [7] used feature engineering and exploratory data analysis to create a feature set for determining the most important factors for predicting the results of a football match, and consequently create a highly accurate predictive system using machine learning. Using gradient boosting they achieved a performance of 0.2156 on the ranked probability score (RPS) metric for game weeks 6 to 38 for the English Premier League aggregated over two seasons (2014–2015 and 2015–2016), whereas the betting organizations that we consider (Bet365 and Pinnacle Sports) obtained an RPS value of 0.2012 for the same period.

IV. PROPOSED MODEL

We present a model to predict the outcome of football matches in the English Premier League. We prepare the dataset of past seasons on various machine learning classifiers. Comparisons amongst the algorithms would be made and the one that turns out to be the most precise i.e. having the sounder forecast accuracy will be considered. Then, optimization can be produced on that classifier to further enhance the model's precision in making forecasts. The tag that would be considered would be Home Win (H), Away Win (A), and Draw (D).

A. Dataset Description

The forecast is done based on data from past games for recent seasons. We have obtained the data set from [8] that has an enormous quantity of data right from the old games to the ones that are being played. There are about 65 features per season like the Home team, Away team, scores, venue to be named a few. After including filtered these features we get about 8-10 features that are going to foresee the outcomes. The dataset size is 6080.

	Date	HomeTear	AwayTean	FTHG	FTAG	FTR	HTGS	ATGS	HTGC	ATGC	HTP	ATP	HM1	HM2	HM3	HM4	HM5	AM1	AM2	AM3	AM4	AM5
2	0	########	Charlton	Man City	4	0	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
3	1	########	Chelsea	West Ham	4	2	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
4	2	########	Coventry	Middlesbr	1	3	NH	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
5	3	########	Derby	Southamp	2	2	NH	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
6	4	########	Leeds	Everton	2	0	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
7	5	########	Leicester	Aston Villa	0	0	NH	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
8	6	########	Liverpool	Bradford	1	0	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
9	7	########	Sunderlan	Arsenal	1	0	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
10	8	########	Tottenhan	Ipswich	3	1	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
11	9	########	Man Unite	Newcastle	2	0	H	0	0	0	0	0	0	M	M	M	M	M	M	M	M	M	M
12	10	########	Arsenal	Liverpool	2	0	H	0	1	1	0	0	1.5	L	M	M	M	M	W	M	M	M	M
13	11	########	Bradford	Chelsea	2	0	H	0	4	1	2	0	1.5	L	M	M	M	M	W	M	M	M	M
14	12	########	Ipswich	Man Unite	1	1	NH	1	2	3	0	0	1.5	L	M	M	M	M	W	M	M	M	M
15	13	########	Middlesbr	Tottenhan	1	1	NH	3	3	1	1	1.5	1.5	W	M	M	M	M	W	M	M	M	M
16	14	########	Everton	Charlton	3	0	H	0	4	2	0	0	1.5	L	M	M	M	M	W	M	M	M	M
17	15	########	Man City	Sunderlan	4	2	H	0	1	4	0	0	1.5	L	M	M	M	M	W	M	M	M	M
18	16	########	Newcastle	Derby	3	2	H	0	2	2	2	0	0.5	L	M	M	M	M	D	M	M	M	M
19	17	########	Southamp	Coventry	1	2	NH	2	1	2	3	0.5	0	D	M	M	M	M	L	M	M	M	M
20	18	########	West Ham	Leicester	0	1	NH	2	0	4	0	0	0.5	L	M	M	M	M	D	M	M	M	M
21	19	########	Arsenal	Charlton	5	3	H	0	4	1	0	0	1.5	L	M	M	M	M	W	M	M	M	M
22	20	########	Bradford	Leicester	0	0	NH	2	1	1	0	1	1.333333	W	L	M	M	M	W	D	M	M	M
23	21	########	Everton	Derby	2	2	NH	3	4	2	5	1	0.333333	W	L	M	M	M	L	D	M	M	M
24	22	########	Ipswich	Sunderlan	1	0	H	2	3	4	4	0.333333	1	D	L	M	M	M	L	W	M	M	M
25	23	########	Man City	Coventry	1	2	NH	4	3	6	4	1	1	W	L	M	M	M	W	L	M	M	M
26	24	########	Middlesbr	Leeds	1	2	NH	4	4	2	1	1.333333	2	D	W	M	M	M	W	W	M	M	M
27	25	########	Newcastle	Tottenhan	2	0	H	3	4	4	2	1	1.333333	W	L	M	M	M	D	W	M	M	M

Fig. 1. Dataset

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11489

B. Data Preprocessing

The dataset that is obtained consists of several attributes of each season. Some of those features are less significant or rather nonessential for foreseeing the outcome. So data cleaning is taken out for keeping only those features that are suitable for the forecast. We have also considered converting categorical data to dummy variables as per requirement.

	FTR	HTP	ATP	HM1	HM2	HM3	AM1	AM2	AM3	HTGD	ATGD	DiffFormPts	DiffLP
30	H	1.25	1.00	D	D	W	D	W	L	0.50	0.25	0.25	-16.0
31	NH	0.75	0.25	L	L	W	D	L	L	-0.50	-0.75	0.50	-2.0
32	H	1.00	1.00	L	D	W	D	W	L	0.00	0.25	0.00	-3.0
33	NH	0.75	0.50	L	L	W	D	L	D	-0.25	-0.25	0.25	3.0
34	NH	1.00	1.50	D	L	W	W	W	L	0.00	0.75	-0.50	3.0

Fig.2. Final Dataset after pre-processing

The following parameters are being considered in our model.
Div = League Division
Date = Match Date (dd/mm/yy)
HomeTeam = Home Team
AwayTeam = Away Team
FTHG = Full Time Home Team Goals
FTAG = Full Time Away Team Goals
FTR = Full Time Result (H=Home Win, D=Draw, A=Away Win)
HTHG = Half Time Home Team Goals
HTAG = Half Time Away Team Goals
HTR = Half Time Result (H=Home Win, D=Draw, A=Away Win)
HTGD = Home Team Goal Difference
ATGD = Away Team Goal Difference
HTP = Home Team Points
ATP = Away Team Points
HM = Home Match
AM = Away Match

C. Exploratory Analysis

In our initial exploratory analysis we found out that the home team clearly has an upper edge over the away team.

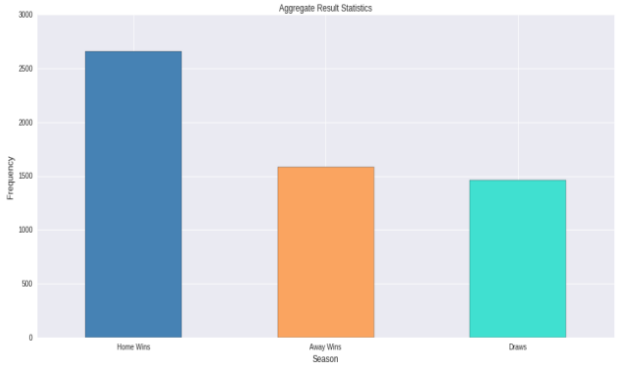

Fig. 3. Distribution of match wins from 2000-2016

Further it was observed that head to head record alone is not a very reliable factor in predicting the end result of a soccer game. Combining other factors such as league points and half time result do make a meaningful impact on end result.

D. Modelling

In this system, we have implemented the following three algorithms: XGBoost, Logistic Regression, and Support Vector Machine (SVM).

Logistic Regression:

Logistic Regression is a Machine Learning method that is used to solve classification issues. It is a predictive analytic technique that is based on the probability idea. The classification algorithm Logistic Regression is used to predict the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11489

likelihood of a categorical dependent variable. The dependent variable in logistic regression is a binary variable with data coded as 1 (yes, True, normal, success, etc.) or 0 (no, False, abnormal, failure, etc.).

Support Vector Machine (SVM):

Support Vector Machines are Machine Learning models which are useful for regression analysis and classification tasks. It falls under the supervised learning category of Machine Learning. These are widely used in classification tasks. Support Vector Machines are based on the idea of finding the best hyperplane that divides the dataset into two parts.

XGBoost:

The XGBoost stands for extreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from gradient boosting. The 'xgboost' is an open-source library that provides machine learning algorithms under gradient boosting methods. The boost.XGBClassifier is a sci-kit-learn API compatible class for classification.

V. EXPERIMENT

An investigation is performed for obtaining the best precision. In this paper, we are operating the data from past recent seasons of the English Premier League. It is done to determine whether the amount of training data has any impact on forecast precision. Following are the accuracy of each model.

Logistic Regression:
F1 score and accuracy for training set: 0.6246 and 0.6654
F1 score and accuracy for test set: 0.6957 and 0.7200.

SVM:
F1 score and accuracy for training set: 0.6470 and 0.6978
F1 score and accuracy for test set: 0.7234 and 0.7400.

XGBClassifier:
F1 score and accuracy for training set: 0.6470 and 0.6978
F1 score and accuracy for test set: 0.7234 and 0.7400.

XGBClassifier with GridSearchCV:
F1 score and accuracy for training set: 0.6318 and0.6777
F1 score and accuracy for test set: 0.7234 and 0.7400.

The results were not only less accurate but they also showed overfitting as we can clearly observe from the accuracy score of training and testing data. In order to get better results ensemble model was required. We studied various other literatures and work and found out that ensemble techniques have been used up in the past giving out an accuracy of about 75%. Since we are already using Xgboost which is an ensemble machine learning algorithm based on gradient boosting,the hyperparameters were tuned by experimenting. The accuracy clearly increased but the overfitting was not resolved. Keeping in mind the variance bias tradeoff, we achieved an accuracy of about **76%**. Also our model is more reliable as we have used 5 fold cross-validation on the same.

VI. RESULT ANALYSIS

Our examination is to foresee the result of the match and when integrated with training data, the XGboost combined with gridsearch cv not only gave better accuracy but also the model is more reliable due to 5 fold cross-validation. So, we have formalized the dataset during the pre-processing phase. Normalization is done to bring the attributes of the training dataset to the same scale. The goal of normalization is to vary the values of numeric columns in the dataset to an ordinary scale, without warping the disparities in the ranges of values [11]. Our model is achieving a f1 score of 76.05% and accuracy of 78.59% on training data and f1 score of 75.00% and accuracy of 76.00% on testing data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 4, April 2022

DOI: 10.17148/IJARCCE.2022.11489

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=0.8,
              enable_categorical=False, gamma=0.4, gpu_id=-1,
              importance_type=None, interaction_constraints='',
              learning_rate=0.06, max_delta_step=0, max_depth=4,
              min_child_weight=5, missing=nan, monotone_constraints='()',
              n_estimators=450, n_jobs=8, num_parallel_tree=1, predictor='auto',
              random_state=2, reg_alpha=1e-05, reg_lambda=1, scale_pos_weight=1,
              seed=2, subsample=0.8, tree_method='exact', validate_parameters=1,
              verbosity=None)
Made predictions in 0.0156 seconds.
F1 score and accuracy score for training set: 0.7605 , 0.7859.
Made predictions in 0.0000 seconds.
F1 score and accuracy score for test set: 0.7500 , 0.7600.
```

Fig. 4. Results after hyperparameter tuning.

VII. CONCLUSION

Sports Analytics is a rapidly growing field and with the advancement of machine learning algorithms, machine learning can be utilized in this domain also. Our aim was to create a model which could effectively predict the result of a soccer game which can then be utilized in various fields like performance analysis, betting industry and fantasy leagues. The game of football not only depends on numbers but also a lot on players and other factors. Also football is a unpredictable sport and combined with the fact that games are usually low scoring, expecting a very high accuracy is not possible. This research can further be improved by taking into account other factors such as a players health statistic or sentiment analysis from twitter.

REFERENCES

[1] Thamaraimanalan, T. Naveena, D. Ramya, m. and Madhubala m., prediction and classification of fouls in soccer game using deep learning. Irish interdisciplinary journal of science & research, 4(3), pp.66-78, 2020.

[2] Yang, predict soccer match outcome based on player performance, Francis academic press, UK, ISSN 2618-1576 vol. 3, issue 3: 74-78, DOI: 10.25236/fsr.2021.

[3] Martinovich, J., Snásel, V., Ochodkova, E., Nolta, L., Wu, J. and Abraham, A., robot soccer-strategy description and game analysis. in ecms (pp. 265-270), June, 2010.

[4] Uhrín Matej, Šourek Gustav, Hubáček Ondřej, Železný Filip, Optimal Sports Betting Strategies In Practice: An Experimental Review, IMA Journal Of Management Mathematics, 10.1093/IMAMAN/DPAA029, (2021).

[5] Woo-Joo Lee, Hyo-Jin Jhang, Seung Hoe Choi, Fuzzy Study On The Winning Rate Of Football Game Betting, Advances In Technology Innovation, 10.46604/AITI.2021.6517, 6, 3, (169-178), (2021).

[6] Edward Wheatcroft, Forecasting Football matches by predicting match statistics, Journal Of Sports Analytics, 10.3233/JSA-200462, (1-21), (2021).

[7] Rahul Baboota, Harleen Kaur, Predictive Analysis And Modeling Football Results Using Machine Learning Approach For English Premier League, International Journal of Forecasting, 10.1016/J.IJFORECAST.2018.01.003, 35, 2, (741-755), (2019).

[8] Rahman, M. M., Faruque Shamim, M. O., & Ismail, S. An Analysis of Bangladesh One Day International Cricket Data: A Machine Learning Approach. International Conference on Innovations in Science, Engineering and Technology (ICISET) 2018.

[9] Football-data . (2019). data | football-data, [online] Available at: http://www.football-data.co.uk / [Accessed on 27 Oct. 2021].

[10] Oughali, M. S., Bahloul, M., & El Rahman, S. A. Analysis of NBA Players and Shot Prediction Using Random Forest and XGBoost Models. International Conference on Computer and InformationSciences(ICCIS) 2019.

[11] Anik, A. I., Yeaser, S., Hossain, A. G. M. I., & Chakrabarty, A. Player's Performance Prediction in ODI Cricket Using Machine Learning Algorithms. 4th International Conference on Electrical 33 Engineering and Information & Communication Technology(iCEEiCT) 2018.

[12] Thirumalai, C., Kanimozhi, R., & Vaishnavi, B. Data analysis using box plot on electricity consumption. International Conference of Electronics, Communication and Aerospace Technology(ICECA) 2017.

[13] statisticshowto.datasciencecentral. (2015). Normalized | statisticshowto.datasciencecentral. [online] Available at: https://www.statisticshowto.datasciencecentral.com/normalized / [Accessed on 27 Oct. 2021].

[14] likegeeks. (2019). Seaborn-heatmap-tutorial | likegeeks. [online] Available at: https://likegeeks.com/seaborn-heatmap-tutorial/ [Accessed on 15 Jan 2022].

[15] towardsdatascience. (2019). Hyperparameter1Tuningc5619e7e6624 | towardsdatascience. [online] Available at: https://towardsdatascience.com/hyperparameter-tuning1c5619e7e6624 [Accessed on 27 Oct. 2021].